DOCUMENT RESUME

ED 424 266 TM 029 139

AUTHOR van der Linden, Wim J.; Scrams, David J.; Schnipke, Deborah L.

TITLE Using Response-Time Constraints in Item Selection To Control for Differential Speededness in Computerized Adaptive Testing. Research Report 98-06.

INSTITUTION Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.

PUB DATE 1998-00-00

NOTE 35p.

AVAILABLE FROM Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.

PUB TYPE Opinion Papers (120) -- Reports - Research (143)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *Adaptive Testing; *Algorithms; *Computer Assisted Testing; Foreign Countries; Linear Programming; Models; Responses; *Selection; *Test Items; Testing Problems; Time

IDENTIFIERS *Constraints; *Speededness (Tests)

ABSTRACT

An item-selection algorithm to neutralize the differential effects of time limits on scores on computerized adaptive tests is proposed. The method is based on a statistical model for the response-time distributions of the examinees on items in the pool that is updated each time a new item has been administered. Predictions from the model are used as constraints in a 0-1 linear programming (LP) model for constrained adaptive testing that maximizes the accuracy of the ability estimator. The method is demonstrated empirically using an item pool from the Armed Services Vocational Aptitude Battery and the responses of 38,357 examinees. The empirical example suggests that the algorithm is able to reduce the speededness of the test for the examinees who otherwise would have suffered from the time limit. Also, the algorithm did not seem to introduce any differential effects on the statistical properties of the theta estimator. (Contains 9 figures and 14 references.) (SLD)

Using Response-Time Constraints in Item Selection to Control for Differential Speededness in Computerized Adaptive Testing

Research Report 98-06

Wim J. van der Linden, University of Twente
David J. Scrams, Educational Testing Service
Deborah L. Schnipke, Law School Admission Council

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

Using Response-Time Constraints in Item Selection to Control for Differential Speededness in Computerized Adaptive Testing

Wim J. van der Linden
University of Twente
David J. Scrams
Educational Testing Service
Deborah L. Schnipke
Law School Admission Council

Abstract

An item-selection algorithm to neutralize the differential effects of time limits on scores on computerized adaptive tests is proposed. The method is based on a statistical model for the response-time distributions of the examinees on the items in the pool that is updated each time a new item has been administered. Predictions from the model are used as constraints in a 0-1 linear programming (LP) model for constrained adaptive testing that maximizes the accuracy of the ability estimator. The method is demonstrated empirically using an item pool from the Armed Services Vocational Aptitude Battery (ASVAB).

Using Response-Time Constraints to Control for Differential Speededness in Computerized Adaptive Testing

Examinees with the same ability differ in the amount of time they need to complete a test item. As a consequence, some examinees may be affected unfavorably by the presence of a time limit and not reach the end of the test. Also, the question of how to score unreached items involves a series of difficult problems. First of all, in a conventional paper-and-pencil test it is impossible to discriminate between unreached items and reached items that were left unanswered because their answers were unknown. But even if it were exactly known which items are not reached (as is possible in computerized testing), scoring would remain a complicated issue. If the test is unspeeded and not too difficult, items not reached might be viewed as missing at random and ignored when the test is scored (for the notion of data missing at random, see Gelman, Carlin, Stern, & Rubin, 1995, sect. 7.4). However, if the test is speeded, the examinees are faced with a speed-accuracy tradeoff and may choose different strategies of responding to the items. The test then needs to be scored under a model that explains such choices, but realistic models for doing so are not yet available.

Empirical studies of the relation between response time and ability have become possible through the introduction of computerized testing but are still hard to find. A favorable exception is a recent study of the response times in a field test of a computerized version of the National Board of Medical Examiners (NBME) Step 2 Licensure Exam (Swanson, Featherman, Case, Luecht, & Nungester, 1997, April). In this study, responses to items in linear subtests were timed, and no correlation between response time and ability was found for the various subtests. However, a replication of the study for the NBME Step 1 Licensure Exam showed moderate correlation towards the end of the subtests (Swanson, personal communication, December 18, 1997). As the Step 1 Exam had a more stringent time limit, the results seem to suggest that for a test with a mild or ineffective time limit, ability and response time are uncorrelated but that a positive correlation is induced if the time limit is tightened.

It seems important to control tests for speededness. Such control would not only make the assumption of a traditional (unidimensional) logistic item response theory (IRT) model more realistic but also prevent scoring problems due to unreached items. However, for a conventional linear test, the only two options seem to be to reduce the length of the test or increase the amount of time available. Given the variability in response times between

examinees, the former would imply loss of accuracy in ability estimation for the faster examinees and the latter an increase in administration costs for all examinees.

In computerized adaptive testing (CAT), an attractive solution to this test design dilemma is possible. If a model for the response-time distributions of the examinees on the items in the pool is available, the actual response times on the items administered to the current examinee can be recorded and used to update the estimates of these distributions for the remaining items in the pool. These distributions can then be used to constrain the selection of the next items in the test to give the test the same degree of speededness for all examinees. Of course, this procedure is only feasible if a model for the response-time distributions with a satisfactory fit to actual response-time data is available.

It is the purpose of this paper to present an algorithm for adaptive testing that builds on this idea. In addition to the usual update of the estimate of the ability parameter in an IRT model, a lognormal model for the response-time distributions is used to update the response-time estimates for the items in the pool. Response-time constraints are derived from these estimates and imposed on the item selection using a 0-1 linear programming (LP) algorithm for constrained adaptive testing. The following sections of this paper introduce the model and the algorithm. The algorithm is then studied empirically using an item pool and estimates of response-time parameters for an adaptive version of a test from the Armed Services Vocational Aptitude Battery (ASVAB). The main purpose of the study was to ascertain the effects of the response-time constraints on the statistical properties of the ability estimator as well as the actual times needed to complete the CAT.

Model for Response Times

The response time of examinee j on item i is denoted by a variable T_{ij}. The variable is assumed to be random because replications of tasks by the same subject are generally known to show variation in the time needed to complete them (Luce, 1986, sect. 1.2; Townsend & Ashby, 1983, chap. 3). The following decomposition for the (natural) logarithm of T_{ij} is assumed as a model for its distribution:

$$\ln T_{ij} = \mu + \delta_i + \tau_j + \varepsilon_{ij}, \qquad (1)$$

with

$$\varepsilon_{ij} \sim N(0,\sigma^2), \tag{2}$$

where μ is the grand mean or general response time level for the item pool and population of examinees, τ_j is an effect parameter for the slowness of examinee j, δ_i for the amount of time demanded by item i, and ε_{ij} is a normally distributed residual or interaction term. Together, (1)-(2) imply a lognormal distribution for the observed response times of a fixed examinee taking a fixed item. The model was proposed in Scrams and Schnipke (in preparation). The effect terms τ_j and δ_i are defined to have expectations equal to zero across examinees and items, respectively. The marginal distribution of log response time across examinees for a fixed item also depends on the distribution of τ. This distribution is examined as one test of the goodness of fit of the model later in this paper.

Observe that the distributions in (1)-(2) vary in location across examinees and items but have a common variance. The last assumption is stringent but allows us to estimate the parameters in the model in a straightforward way. In addition, since the model will be used to constrain item selection using only a percentile in the upper tail of the distributions towards the end of the test, a slight misfit of the model would not seem to lead to serious item selection errors . However, whenever using the model, it should be standard practice to check its assumptions.

For future reference, note that

$$\mu \equiv E_{ij}(\ln T_{ij}), \tag{3}$$

$$\delta_i \equiv E_j(\ln T_{ij}) - \mu, \tag{4}$$

$$\tau_j \equiv E_i(\ln T_{ij}) - \mu, \tag{5}$$

$$\sigma^2 \equiv E_{ij}[\ln T_{ij} - \delta_i - \tau_j]^2. \tag{6}$$

Throughout this paper subscripts at expectation signs denote indices over which expectations are taken.

A different use of the lognormal distribution as a model for response times is made in

Thissen's (1983) model for timed testing. In his model, the lognormal distribution is parameterized to be dependent on the latent ability measured by the items and becomes part of the likelihood function used for estimating the examinees' ability from the joint distribution of the item scores and the response times. Thissen also found adequate fit for a series of tests, except for one which showed an overrepresentation of fast responses due to the relative easiness of its items. Other distributions used to study response times on test items are the Weibull (Roskam, 1997) and the gamma distribution (Verhelst, Verstralen & Jansen, 1997). In a previous study, the lognormal distribution showed a good fit to the response time distributions on an item pool from the Armed Services Vocational Aptitude Battery (ASVAB), outperforming the Weibull and gamma distributions (Schnipke & Scrams, 1997). These results will be further discussed below when an empirical example for the ASVAB is presented.

IRT Model

It is assumed that the item pool has been calibrated using an IRT model. In the empirical example later in this paper, the item pool was calibrated using the 3-parameter logistic (3-PL) model. The model describes the probability of a correct response on item i as:

$$p_i(\theta) \equiv \text{Prob}\{U_i = 1 | \theta\} \equiv c_i + (1 - c_i)\{1 + \exp[-a_i(\theta - b_i)]\}^{-1}, \quad (7)$$

where θ is the unknown ability of the examinee and $a_i \in [0,\infty]$, $b_i \in [-\infty,\infty]$, and $c_i \in [0,1]$ are the discrimination, difficulty, and guessing parameter for item i, respectively (Lord, 1980, chap. 2).

A key quantity in IRT is Fisher's information on the unknown ability parameter. For a test of n items, the measure is defined as:

$$I_{U_1,\ldots,U_n} \equiv -E\left(\frac{\partial}{\partial \theta^2} \ln L(\theta | U_1,\ldots,U_n)\right). \quad (8)$$

In (8), $L(\theta | U_1,\ldots,U_N)$ is the likelihood statistic associated with the (random) response vector $U_1,\ldots,U_n$. For the 3-PL model in (7) it holds that

$$I_{U_1,\ldots,U_n}(\theta) = \sum_{i=1}^{n} \frac{(p_i'(\theta))^2}{p_i(\theta)(1-p_i(\theta))}, \tag{9}$$

with

$$p_i'(\theta) \equiv \frac{\partial}{\partial\theta} p_i(\theta) \tag{10}$$

(Lord, 1980, chap. 5). The term in (9) for item i will be denoted as $I_i(\theta)$ and is the item information used in the maximum-information item selection criterion in the CAT algorithm below.

Response-Time Constraints in CAT

It is assumed that the item pool consists of items indexed by i=1,...,I. In addition, the CAT is assumed to consist of items indexed by k=1,...,n. Thus, index i_k represents the event of the ith item in the pool being administered as the kth item in the CAT. The index values of the first k-1 items in the CAT are denoted by the set $S_{k-1} \equiv \{i_1,\ldots,i_{k-1}\}$. The remaining items in the pool are denoted by the set $R_k \equiv \{1,\ldots,I\} \backslash S_{k-1}$. The kth item in the test is chosen from the set R_k.

The basic idea is to update an estimate of the examinee's slowness parameter τ_j in (1)-(2) during the test given accurate estimates of μ, item parameters δ_i, i=1,...,I, and the residual variance, σ^2. The improved estimates of τ_j are used to update projections of the time needed to complete each of the remaining items in the pool. The next item is then selected subject to a constraint based on these projections as well as the time available to complete the remaining portion of the test. A Bayesian framework is used to update the response time projections whereas the response time constraints are incorporated in the item selection procedure using a 0-1 linear programming (LP) model for constrained CAT that maximizes the information on θ in the test and also allows for additional constraints that can be used to guarantee its content validity.

Updating Response Time Estimates

It is assumed that δ_i, i=1,...,I, and σ^2 have been estimated precisely enough to be considered as known. Estimates can easily be obtained from the response times in the calibration sample using the equations in (4) and (6). However, if examinee j is tested, τ_j is an unknown parameter; it is assumed to have a normal prior distribution:

$$\tau_j \sim N(\mu_{0j}, \sigma^2_{0j}). \tag{11}$$

The model in (1)-(2) yields a normal likelihood with unknown mean and known variance that has the family of normal distributions as its conjugate prior (Gelman, Carlin, Stern & Rubin, 1995, sect. 2.6). Hence, noting $\tau_j \equiv \ln(t_{ij}) - \mu - \delta_i + \varepsilon_{ij}$, the posterior distribution of τ_j, after the response times on items $i_1,...,i_{k-1}$ have been recorded, is normal with mean and variance:

$$E(\tau_j \mid t_{i_1 j},...,t_{i_{k-1} j}) = [\sigma^2 \mu_{0j} + \sigma^2_{0j} \sum_{p=1}^{k-1} (\ln(t_{i_p j}) - \mu - \delta_{i_p})] / [\sigma^2 + (k-1)\sigma^2_{0j}] \tag{12}$$

$$Var(\tau_j \mid t_{i_1 j},...,t_{i_{k-1} j}) = \sigma^2_{0j}\sigma^2 / ((k-1)\sigma^2_{0j} + \sigma^2) \tag{13}$$

Also, the predictive density for the response time of examinee j on item i after items $i_1,...,i_{k-1}$ is normal with mean equal to the posterior mean and variance equal to the sum of the prior and posterior variances, respectively:

$$E(\ln T_{ij} \mid t_{i_1 j},...,t_{i_{k-1} j}) = E(\tau_j \mid t_{i_p j},...,t_{i_{k-1} j}) + \mu + \delta_i \tag{14}$$

$$Var(\ln T_{ij} \mid t_{i_1 j},...,t_{i_{k-1} j}) = \sigma^2_{0j} + Var(\tau_j \mid t_{i_p j},...,t_{i_{k-1} j}). \tag{15}$$

As the examinees are assumed to be exchangeable, an obvious choice for the parameters in this prior is to equate them to the mean and variance of the population of examinees:

$$\mu_{0j} \equiv E(\tau) = E_j E_{i|j}(\ln T_{ij}) = 0, \tag{16}$$

$$\sigma^2_{0j} \equiv \sigma^2_\tau. \tag{17}$$

for all j. Using (16)-(17), the mean and variance in (14)-(15) specialize to:

$$E(\ln T_{ij} | t_{i_1 j},\ldots,t_{i_{k-1} j}) = \mu + \delta_i + \frac{\sum_{p=1}^{k-1} (\ln(t_{i_p j}) - \mu - \delta_{i_p})}{\sigma^2/\sigma^2_\tau + k - 1} \tag{18}$$

$$Var(\ln T_{ij} | t_{i_1 j},\ldots,t_{i_{k-1} j}) = \frac{\sigma^2 + k\sigma^2_\tau}{1 + (k-1)\sigma^2_\tau/\sigma^2}. \tag{19}$$

Because δ_i, i=1,...,I, and σ^2 are assumed to be estimated using (4) and (6), respectively, the expressions in (18) and (19) have only known constants and are easy to calculate.

Let t_{ij}^{α} be the αth certain percentile in this posterior predictive density for $\ln(T_{ij})$ transformed back to the original time scale. The choice of item i_k will now be constrained using this percentile for all remaining items in the pool, $i \in R_k$. As will become clear later, it makes sense to define α to be dependent on k and choose a percentile near the middle of the density in the beginning of the test and move to percentiles in the upper tail towards the end of the test.

Constrained CAT Algorithm

The kth item is selected according to an algorithm for constrained CAT presented in van der Linden and Reese (1998). To select the initial item, the algorithm first selects a full test that meets all constraints to be imposed on the selection of items in the CAT and has maximum information at the initial ability estimate. The item actually administered is the one from the assembled test with maximum information at this ability estimate. At each next step, the test is reassembled to have maximum information at the updated ability estimate fixing the items already administered. Again, the item to be administered is selected from the new portion of the test to have maximum information at the updated ability estimate. The procedure is repeated until the last item is selected.

The fact that a full test is assembled at each step rather than a single item is to keep future item selection feasible with respect to the set of constraints. Because both the test assembly and the selection of the individual item have the objective of maximum information, the ability estimator can be expected to be maximally informative too. All test assembly is done while the examinee takes the test and is based on a 0-1 LP model that represents all test specifications.

To discuss the model, decision variables x_i, i=1,...,I, are introduced that take the value 1 if item i is selected in the test and the value 0 otherwise. The total amount of time available for the CAT is denoted as t_{tot}. In addition, it is assumed that the composition of the test is constrained with respect to a variety of categorical attributes, such as content, cognitive level, and item format. These attributes partition the item pool into a collection of sets V_g, g=1,...,G, each of which is defined by a (combination of) attribute value(s). Also, the composition of the test can be constrained with respect to several quantitative attributes a_{hi}, h=1,..,H, for example, word counts, exposure rates, and IRT parameter values. Finally, let $\hat{\theta}_{k-1}$ be the estimate of θ after k-1 items have been administered.

The decision variables are used to formulate the following linear model for selecting item k for the current examinee:

$$\text{maximize} \sum_{i=1}^{I} I_i(\hat{\theta}_{k-1})\, x_i \tag{20}$$

subject to

$$\sum_{i \in S_{k-1}} t_{ij}\, x_i + \sum_{i \in R_k} t_{ij}^{\alpha k}\, x_i \leq t_{tot}, \tag{21}$$

$$\sum_{i \in S_{k-1}} x_i = k-1, \tag{22}$$

$$\sum_{i=1}^{I} x_i = n \tag{23}$$

$$\sum_{i \in V_g} x_i \geq n_g^{(1)}, \quad g=1,...,G, \tag{24}$$

$$\sum_{i \in V_g} x_i < n_g^{(2)}, \quad g=1,...,G, \tag{25}$$

$$\sum_{i=1}^{I} a_{hi} x_i \geq n_h^{(1)}, \quad h=1,...,H, \tag{26}$$

$$\sum_{i=1}^{I} a_{hi} x_i < n_h^{(2)}, \quad h=1,...,H, \tag{27}$$

$$x_i \in \{0,1\}, \quad i=1,...,I. \tag{28}$$

The objective function in (20) optimizes the information in the test at $\hat{\theta}_{k-1}$. The total length of the test is set at n items in (23), whereas (22) fixes the values of the decision variables of all items that have already been administered to the examinee at 1. The key constraint in this paper is the one on response times in (21) which requires the remaining n-k+1 items to be selected such that the sum of the α_k th percentiles of their predicted response-time distributions plus the actual response time on the first k-1 items not be larger than the total amount of time available. Note that these percentiles are now defined to be dependent on the rank of the item in the test. In (24)-(25), the numbers of items with the various attribute categories are required to be between lower and upper bounds $n_g^{(1)}$ and $n_g^{(2)}$, respectively. Finally, the constraints in (26)-(27) guarantee that the sums of values for the various quantitative attributes are between the bounds $n_h^{(1)}$ and $n_h^{(2)}$.

At step k, the model thus selects n-k+1 new items from the set R_{k-1}. The item actually administered is the one selected from this set that is most informative at $\hat{\theta}_{k-1}$. The cycle is then repeated to select item k+1.

As already noted, it makes sense to choose $t_{ij}^{\alpha k}$ close to the means of the posterior predictive response time densities in the beginning of the test but to move towards their upper tails later on. This suggestion is motivated by the fact that the sum of the mean predicted response times is a good predictor of the actual time for a large set of items but a more

conservative predictor is needed if the set becomes smaller.

A larger selection of constraints can be added to the model to deal with possibly remaining test specifications than the set used in (20)-(28). Several examples of possible additions are given in van der Linden (1998) and van der Linden and Reese (1998). The set of specifications should be large enough to guarantee a CAT with satisfactory content validity.

Empirical Example

A previous pool from the Armed Services Vocational Aptitude Battery (ASVAB) was used to study the behavior of the algorithm. The CAT version of the ASVAB is extensively described in Sands, Waters and McBride (1997). The pool consisted of 186 items used for the Arithmetic Reasoning Test. The length of the test was 15 items. The items in the pool were calibrated using the 3-PL model in (7). A dominant feature of the pool was its high positive correlation between the estimated values of the difficulty and discrimination parameters displayed in Figure 1. This correlation explains an unexpected result reported below.

[Figure 1 about here]

Response-time data were recorded for 38,357 examinees who took the test in 1997. The parameters in the response-time model in (1)-(2) were estimated substituting sample statistics into (3)-(6). The following results were obtained: μ=4.093, σ=.515. In addition, the estimated item and person effects, δ_i and τ_j, were found to be distributed with a standard deviation equal to .497 and .375, respectively. As shown in Figure 2, the values of the estimated item effect

[Figure 2 about here]

parameter, δ_i, increased strongly with the estimated difficulty of the items. This result is as expected for an arithmetic reasoning test--the more difficult the item, the more timed needed to solve it. Also, the correlation between θ and τ was found to be equal to .035, indicating that ability and speed were independent variables in the population of examinees.

Note that the sample standard deviations of and τ reported here no doubt overestimate their population parameters. However, because the sample of examinees is large, the bias in the estimate of the standard deviation of δ will be negligible. Moreover, this estimate is only reported here as a descriptive statistic; it does not play any role in the adaptive procedure in this example. The bias in the estimate of σ_τ is expected to be larger but this quantity serves as the variance of the prior for τ_j in the adaptive procedure. The result is thus a less informative prior,

and hence a more conservative adaptive test.

Also, note that the matrix of the CAT-ASVAB response times used in this example was not necessarily balanced with respect to the person and item effects. It is hard to ascertain how this has affected the estimated item effects and the standard deviation of the examinee effects. However, serious effects would have led to a deteriorated fit of the lognormal model. Since the model fit was general good (see below), confounding was not considered an important issue. For an actual application of the lognormal model, it is recommended to collect response time data through a balanced design (e.g., during item pretesting) or to make the matrix balanced afterwards through resampling. The latter option was not feasible here because for each examinee response times on only 15 items were available.

The main goal of the study was to explore the effects of the response-time constraints in (21) on the statistical properties of the ability estimator. In particular, the bias and root mean squared error (RMSE) functions of the estimator were studied relative to those for the unconstrained version of the CAT-ASVAB. In an earlier study of constrained adaptive testing for an item pool from the Law School Admission Test (LSAT) with an 0-1 LP model with 433 constraints representing its specifications, the bias and RMSE functions were slightly affected by the presence of the constraints for short test lengths but no effects remained after 30 items (van der Linden & Reese, 1998).

A second goal of the study was to assess the effect of the use of the response-time constraints on the examinees' time needed to complete the test. In the main simulation study, the time limit was set equal to the one was used in the actual ASVAB test (t_{tot}=39 mins). However, the limit was chosen to introduce only a mild form of speededness for the ASVAB examinees (Segall, Moreno, Bloxom, & Hetter, 1997). To investigate the effects of tighter time limits, this part of the simulation was therefore replicated for t_{tot}=34 and 29 mins. Finally, to create a baseline for evaluating the effects of the response-time constraints, the times needed to complete the test for an unconstrained CAT version were also simulated.

The ability values of the simulated examinees were chosen to be equal to θ = -2.0,-1.5,...,2.0. In addition, their response times were simulated at τ = -.60, -.30, .00, .30, and .60. It is reminded that the values of τ in the sample of ASVAB examinees were estimated to be distributed about .00 with a standard deviation equal to .515; the τ values were thus chosen to cover the range of values in the ASVAB population. The number of replications for each combination of the θ and τ values was equal to 180.

As already mentioned, all simulations were replicated without the response-time

constraints on the item selection. This condition provided a base line for the evaluation of the constrained CATs.

The ability estimator used was the expected value of the posterior distribution of the ability parameter (EAP estimator), with a uniform prior distribution over $\theta \in [-4.0,4.0]$. The first item was selected to be optimal at $\theta = 0$. The values of the parameter $t_{ij}^{\alpha k}$ in (21) were set equal to the 50th percentile in the posterior predictive distributions for the selection of item k=1 and moved to the 95th percentile for the selection of item k=13 in equal steps. The last value was maintained for items k=14 and 15.

The LP model was solved using the First Acceptable Integer Solution Algorithm from the ConTEST test assembly software package; a detailed description of the algorithm is available in the manual (Timminga, van der Linden, & Schweizer, 1996, sect. 6.6). On a PC with Pentium/133MHz processor the times needed to update the ability estimate, solve the LP model, and select the most informative item were always less than 1.5 secs.

Fit of Response-Time Distribution

The response-time distribution in (1)-(2) was tested for its assumption of a lognormal shape against the assumptions of a normal, gamma, and Weibull distribution. Detailed results for the current data set are given in Schnipke and Scrams (1997). Only single observations were recorded for each item-examinee combination. However, as shown by Figure 3, the distribution of the values for the examinee slowness parameter, τ_j, approximated a normal

[Figure 3 about here]

distribution and therefore the marginal distribution of the log response times across the examinees in the ASVAB population should also be approximately normal. This feature was checked in depth for 30 of the 186 items. The items were selected to be answered by at least 1,000 examinees and not to involve any figures. The samples were randomly split into halves used for estimating the parameters and checking the distributional assumption. The parameters of the four candidate models for the response-time distributions were estimated using the method of ML estimation.

Double probability plots were produced for each model and each item, with the observed cumulative probability function along the abscissa and the estimated function along the ordinate. A typical example is given in Figure 4. The lognormal distribution provided the

[Figure 4 about here]

best fit (indicated by most points falling along the diagonal), followed by the gamma,

Weibull, and normal. The same essential result was found for all other items.

Fits were also examined using the RMSE calculated between the observed and estimated distribution functions at each fifth percentile. The results for all 30 items are provided in Figure 5. For readability, items were ranked according to the quality of fit provided by the

[Figure 5 about here]

lognormal distribution. Again, the lognormal distribution provided the best overall fit, and the other three distributions were ranked as before.

The assumption of a common standard deviation across items was tested by plotting the sample estimates of the standard deviations of the same marginal distributions of the log response times for each of the 186 items against sample sizes in Figure 6. Different items

[Figure 6 about here]

items were administered different numbers of times. For smaller samples much of the observed variability may be attributed to sampling variation. However, for larger samples, the estimated standard deviations should stabilize about an identifiable mean. Figure 6 shows this stabilization to hold indeed about the value of 0.63.

Bias and RMSE Functions of Estimator of θ

The bias and RMSE of the θ estimator in the CAT algorithm were estimated with results displayed in Figure 7. Bias as a function of θ was roughly flat except for the upper part of the

[Figure 7 about here]

scale, where a larger positive bias was obtained. This effect was observed both for the constrained and unconstrained CAT versions. It is assumed to be entirely due to the distribution of the values for the difficulty and discrimination parameters in the item pool. As is evident from Figure 1, examinees with a θ value at the upper part of the scale will tend to get items that are too easy with an extremely high value for their discrimination parameter. Most of these items will yield correct responses and the ability estimates will tend to drift away. Bias in the estimator of θ as a function of τ was flat for all θ values indicating no systematic impact of the personal slowness parameter on the errors in the θ estimates. Observe that the altitudes of the lines correspond with the average bias for the θ values in Figure 7a.

The RMSE plot in Figure 7c shows the typical U-shaped form for a CAT with an initial ability estimate at θ=0. The lack of asymmetry must be the result of the squared bias component in the mean-squared error originated from the composition of the item pool discussed above. Again, both the constrained and unconstrained CAT versions have this

asymmetry. Figure 7d reveals that the slowness of the examinees had no impact on RMSE of the θ estimator.

A comparison between the bias and RMSE for the constrained and unconstrained CATs revealed slightly unfavorable differences for the constrained function. The differences are assumed to be the result of the relatively short length of 15 items for the adaptive test. In the previous study with the LSAT, larger differences were found for a test length of 15 items; however, these differences disappeared completely as the test length approached 30 items (van der Linden & Reese, 1998).

Bias and RMSE Functions of Estimator of τ

Also, the bias and RMSE of the τ estimator were estimated as a function of θ and τ. The results are given in Figure 8. The plots show a small inward bias, typical of a Bayesian estimator,

[Figure 8 about here]

and a uniform RMSE as a function of τ (Figures 8a and 8c). In addition, as expected, the bias and RMSE appear to be independent of the true value of the θ parameter. As before, the altitudes of the horizontal bias functions in Figure 8b correspond with the bias at the corresponding τ values in Figure 8a.

Distributions of Time Left After Completion

Because response times were sampled from the model in (1)-(2) for each item selected it was possible to estimate for each simulated examinee how much time was needed to complete the test. Figure 9a shows the average time left after completion of the CAT version without

[Figure 9 about here]

response-time constraints as a function of the examinee slowness parameter, τ. The dotted line represents the time limit of 39 mins. (=2340 secs.) in the simulation; results below the line indicate extra time needed to finish the full test. For all θ values, the average time remaining at the end of the CAT is a decreasing function of the slowness of the examinee. The majority of the simulated examinees appeared to complete the test in time; those who were not able to do so were exclusively among the ones with high θ values. In a CAT, these examinees get the most difficult items, and the positive correlation between b_i and δ_i in Figure 2 shows that these items tend to demand more time.

Figure 9b shows the effect of the response-time constraints. The curves for the examinees with the high θ values now run more horizontally for the larger τ values, and none of

the curves intersect the line representing the time limit. As the time limit for the ASVAB was rather mild, the simulations with the response-time constraints were repeated for t_{tot} equal to 34 mins (=2040 secs.) and 29 mins. (=1740 secs.). As Figures 9c and 9d show, the general effect of a tighter time limit is less time after completion, but all curves are still above the line representing the limits.

Conclusion

The purpose of this paper was to present an item-selection algorithm for CAT to remove the differential effects of time limits on the performances of examinees. The empirical example suggests that the algorithm is able to reduce the speededness of the test for the examinees who otherwise would have suffered from the time limit. Also, the algorithm did not seem to introduce any differential effects on the statistical properties of the θ estimator. The differences in bias and RMSE in this estimator between the constrained and unconstrained CAT versions were uniform and independent of any other parameter in the simulation. Also, they were generally small and are expected to disappear for a longer test than the subtest from the ASVAB used in the example.

The examinees in the empirical example who suffered from the time limit under the condition of an unconstrained CAT were thus exclusively among those with high θ values. This finding was a direct result from the fact that ability and speed were independent variables in the examinee population whereas item difficulty and the time demanded by the item correlated positively. As speed and ability were uncorrelated, some of the more able examinees were slow. Nevertheless, as a result from the adaptive nature of the test and the positive correlation between item difficulty and response time, they tended to get the items that demanded most time. These examinees thus profited strongly from the presence of the response-time constraints in the proposed algorithm.

The low impact of the time constraints on the estimator of θ suggests that the time limit used for the CAT-ASVAB was ample. The additional analyses with the 34- and 29-minute time limits suggest that the item pool may be rich enough to support even more stringent time limits. Ultimately, however, there is a tradeoff between any constraint on the item selection and the precision of measurement. Any operational use of this algorithm would therefore require a detailed analyses of the richness of the item pool and the complexity of the other constraints imposed on the item selection.

References

Gelman, A., Carlin, J.B, Stern, H., & Rubin, D.B. (1995). Bayesian data analysis. London: Chapman & Hall.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Luce, D.R. (1986). Response times: Their role in inferring elementary mental organization. New York: Oxford University Press.

Roskam, E.E. (1997). Models for speed and time-limit tests. In W.J. van der Linden & R.K. Hambleton (Eds.), Handbook of modern item response theory (pp. 187-208). New York: Springer-Verlag.

Sands, W.A., Waters, B.K., & McBride, J.R. (Eds.) (1997). Computerized adaptive testing: From inquiry to operation. Washington, D.C.: American Psychological Association.

Schnipke, D.L., Scrams, D.J. (1997). Representing response-time information in item banks (LSAC Computerized Testing Report No. 97-09). Newtown, PA: Law School Admission Council.

Scrams, D.J., & Schnipke, D.L. (in preparation). A lognormal response time model. Princeton, NJ: Educational Testing Service.

Segall, D.O., Moreno, K.E., Bloxom, B.M., Hetter, R.D. (1997). Psychometric procedures for administering CAT-ASVAB. In W.A. Sands, B.K. Waters, & J.R. McBride (Eds.). Computerized adaptive testing: From inquiry to operation. Washington, D.C.: American Psychological Association.

Swanson, D.B., Featherman, C, Case, S.M., Luecht, R.M., & Nungester, R.J. (1997, April). Relation of response latency to test design, examinee proficiency, and item difficulty in computer-based test administration. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Thissen, D. (1983). Timed testing: An approach using item response testing. In D.J. Weiss (Ed.), New horizons in testing: Latent trait theory and computerized adaptive testing (pp. 179-203). New York: Academic Press.

Townsend, J.T., & Ashby, G.F. (1983). The stochastic modeling of elementary psychological processes. Cambridge: Cambridge University Press.

van der Linden, W.J. (1988). Optimal assembly of psychological and educational tests. Applied Psychological Measurement, 22 (in press).

van der Linden, W.J. & Reese, L.M. (1998). An optimal model for constrained adaptive testing. Applied Psychological Measurement, 22 (in press).

Verhelst, N.D., Verstralen, H.H.F.M., & Jansen, M.G.H. (1997). A logistic model for time-limit tests. In W.J. van der Linden & R.K. Hambleton (Eds.), Handbook of modern item response theory (169-186). New York: Springer-Verlag.

Authors' Note

This study received funding from the Law School Admission Council (LSAC). The opinions and conclusions contained in the paper are those of the authors and do not necessarily reflect the position or policy of LSAC. The authors are indebted to Wim M.M. Tielen for writing the software used in the empirical example, Jackelien ter Burg for her computational assistance, and the Defense Manpower Data Center for the permission to use the ASVAB data set.

Figure Captions

Figure 1. Bivariate plot of the estimated values of the discrimination and difficulty parameters in the item pool.

Figure 2. Bivariate plot of the estimated values of the δ and difficulty parameters in the item pool.

Figure 3. Frequency diagram of the estimated values of the examinee slowness parameter with the best fitting normal density curve.

Figure 4. Double probability plot of the four fitted response-time distributions for a typical item.

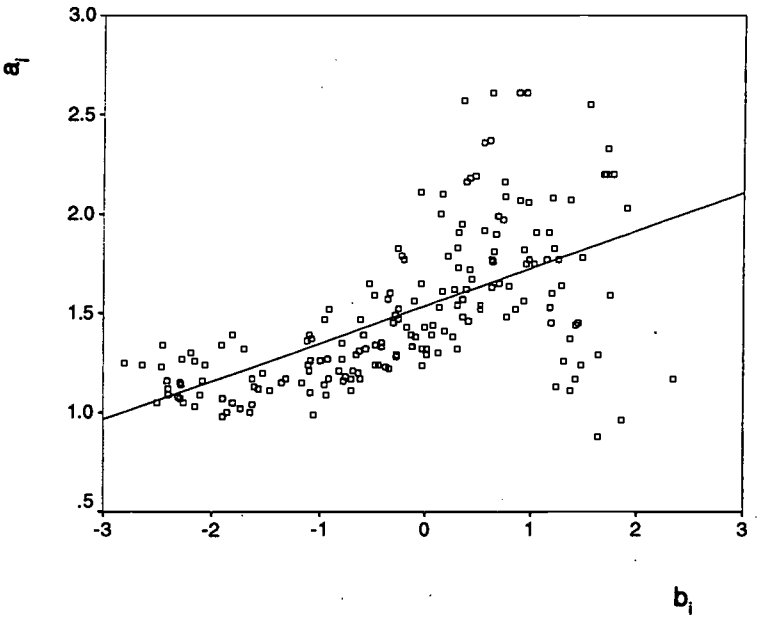
Figure 5. RMSE of the items for the four response time-distributions.

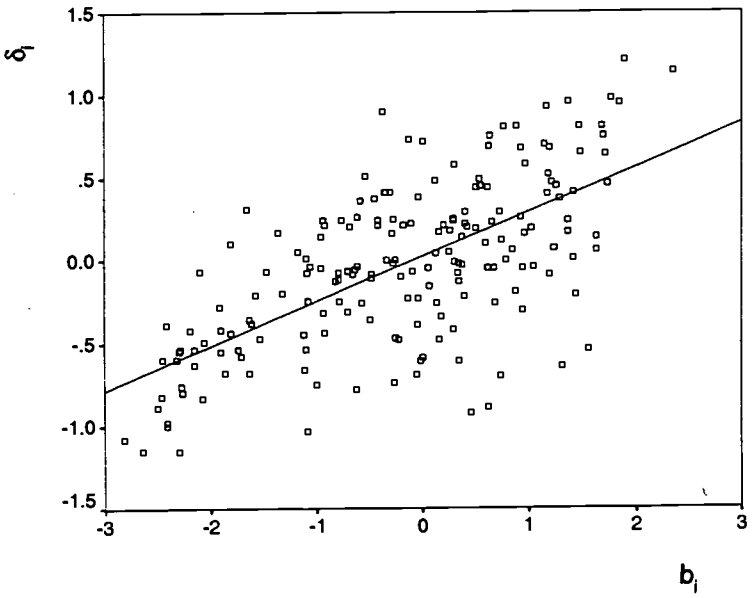
Figure 6. Estimated standard deviations of the marginal log response-time distributions for the items as a function of the size of the examinee sample.

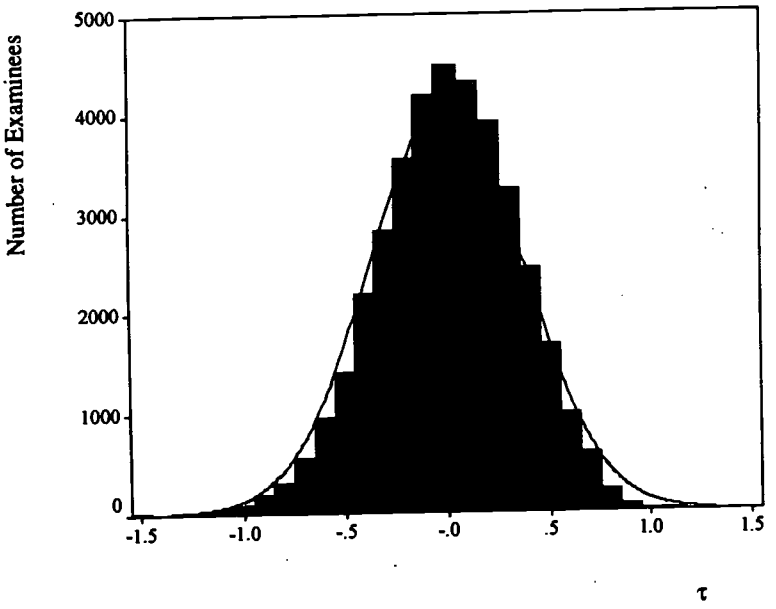
Figure 7. Estimated bias and RMSE of the θ estimator as a function of θ and τ.

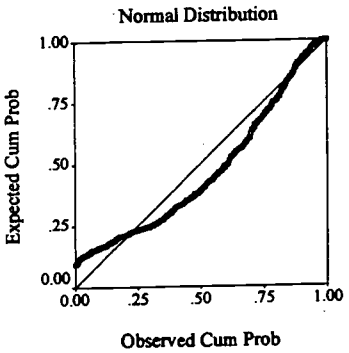
Figure 8. Estimated bias and RMSE of the τ estimator as a function of τ and θ.

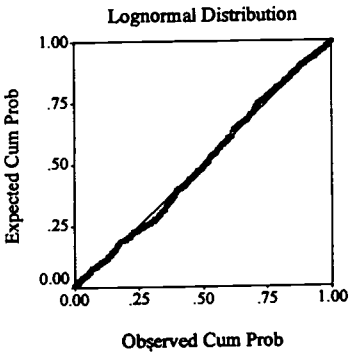
Figure 9. Average time left after completion of the test for CAT without (9a) and with response-time constraints on item selection (9b: t_{tot}=39; 9c: t_{tot}=34: 9d: t_{tot}=29 mins.).

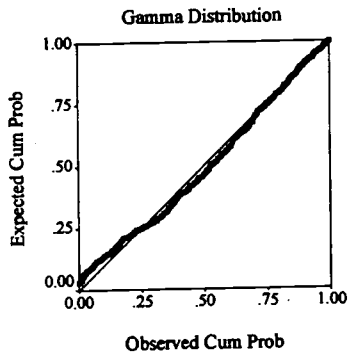

3.0
2.5
2.0
1.5
1.0
.5
a_i
-3
-2
-1
0
1
2
3
b_i

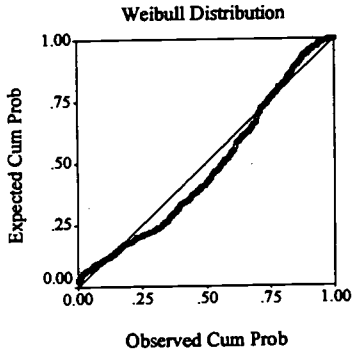

1.5
1.0
.5
0.0
-.5
-1.0
-1.5
-3
-2
-1
0
1
2
3
δ_i
b_i

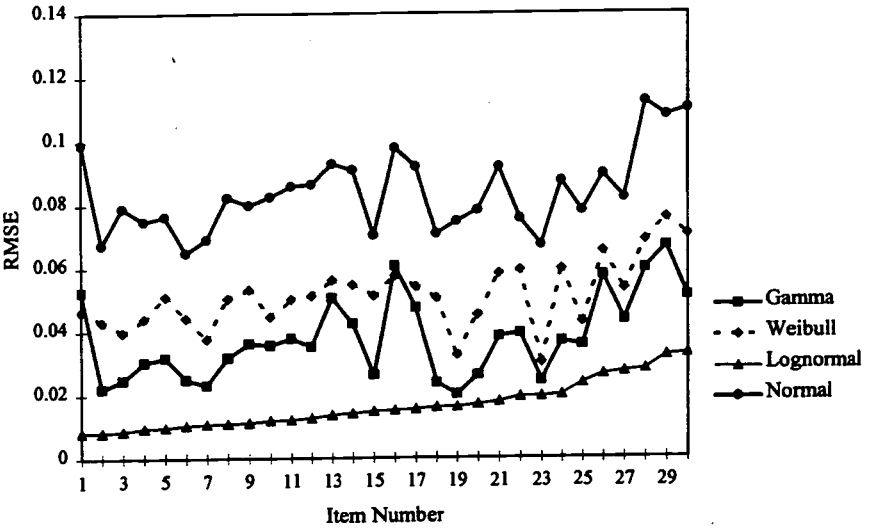

Number of Examinees
5000
4000
3000
2000
1000
0
-1.5
-1.0
-.5
-.0
.5
1.0
1.5
τ

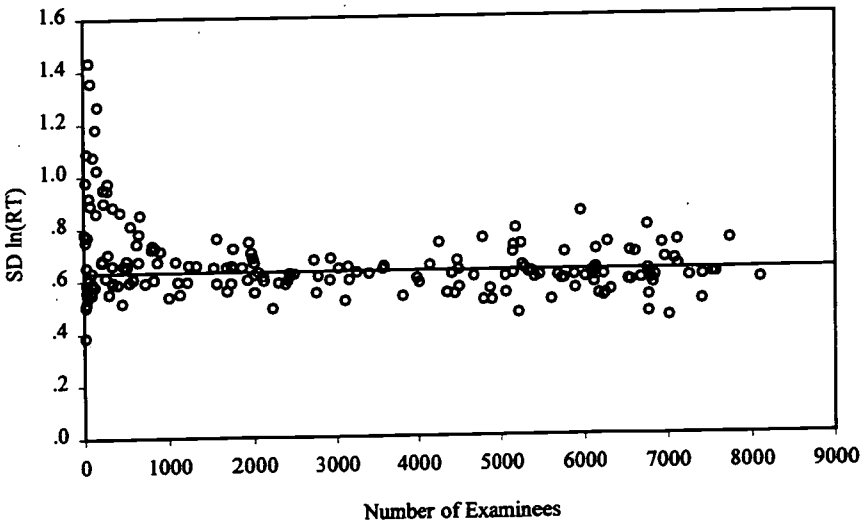

Normal Distribution
Expected Cum Prob
1.00
.75
.50
.25
0.00
0.00
.25
.50
.75
1.00
Observed Cum Prob

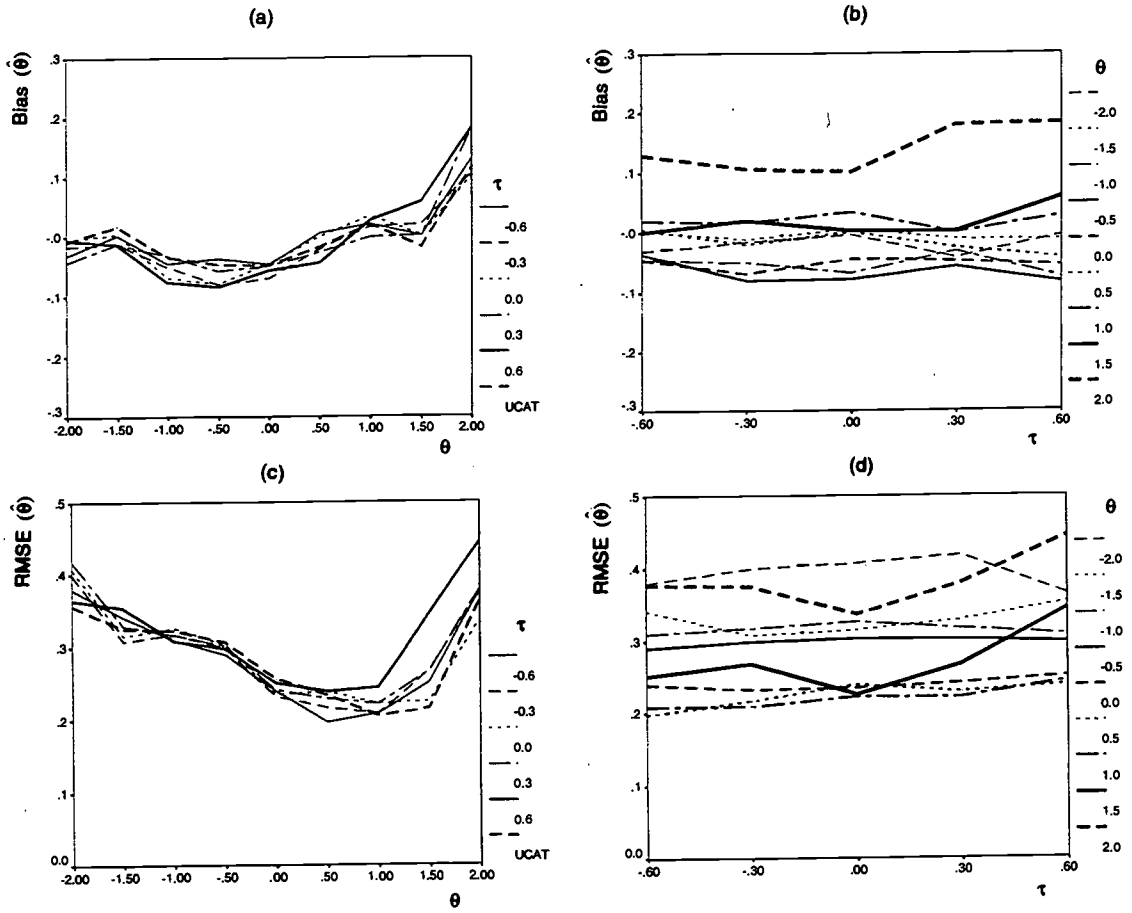

Lognormal Distribution
Expected Cum Prob
1.00
.75
.50
.25
0.00
0.00
.25
.50
.75
1.00
Observed Cum Prob

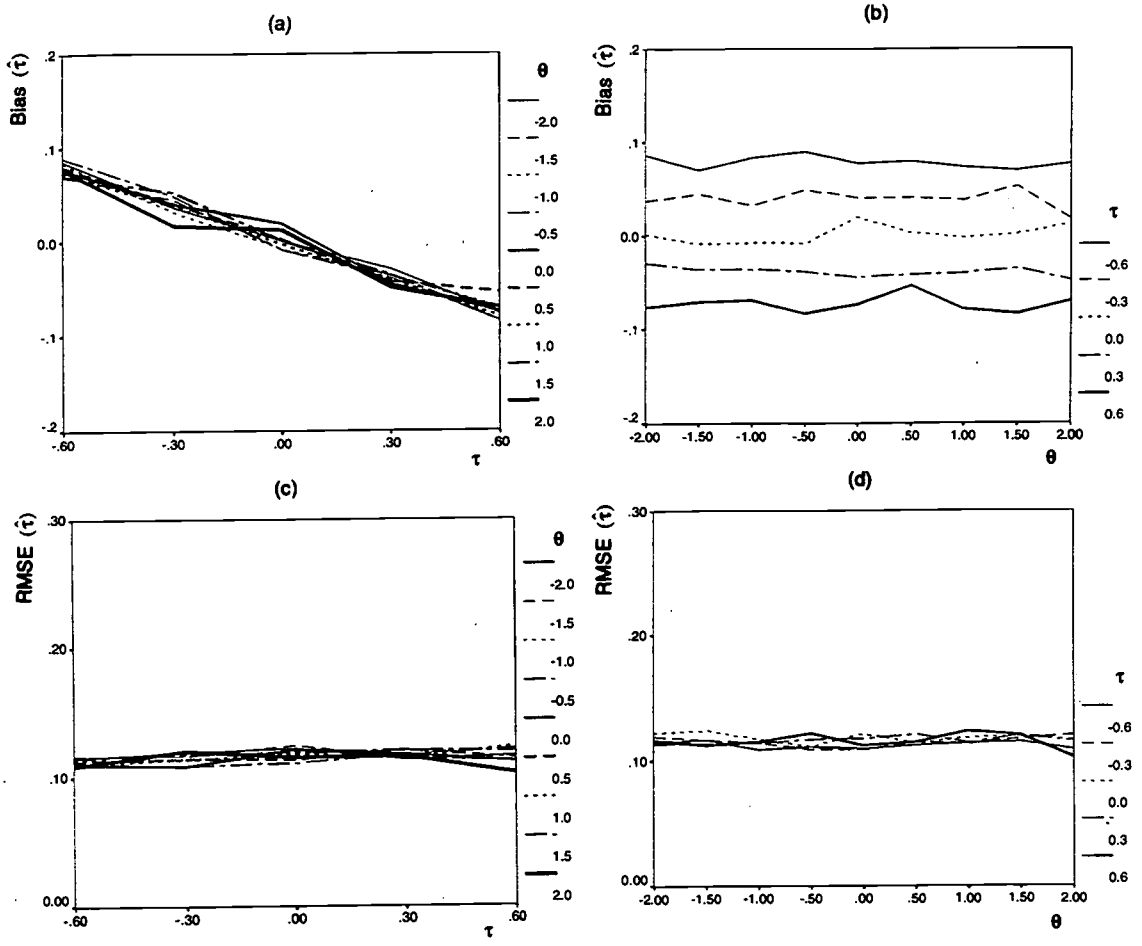

Gamma Distribution
Expected Cum Prob
1.00
.75
.50
.25
0.00
0.00
.25
.50
.75
1.00
Observed Cum Prob

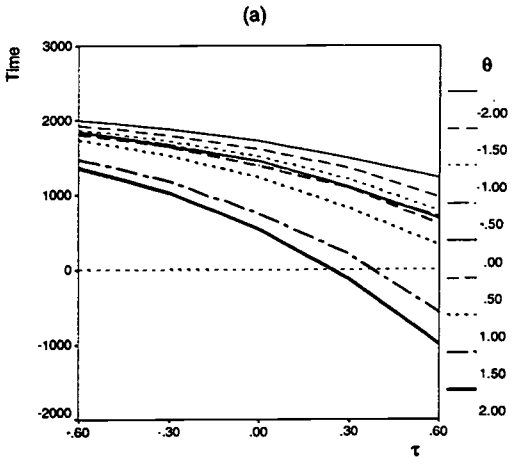

Weibull Distribution
Expected Cum Prob
1.00
.75
.50
.25
0.00
0.00
.25
.50
.75
1.00
Observed Cum Prob

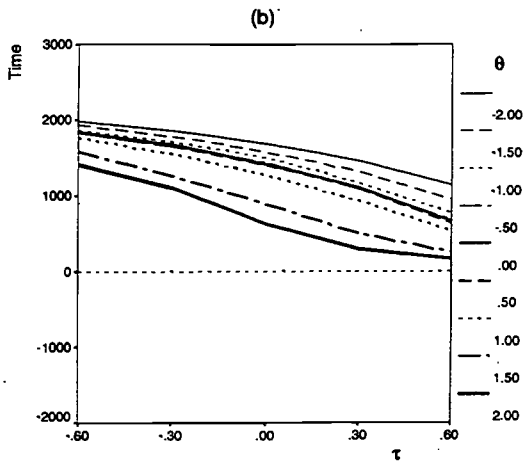

0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
RMSE
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
Item Number
Gamma
Weibull
Lognormal
Normal

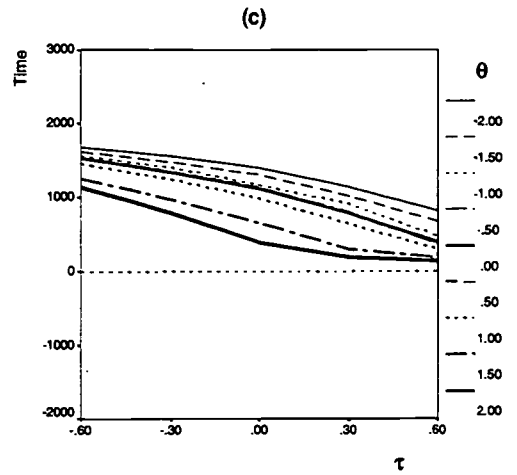

1.6
1.4
1.2
1.0
.8
.6
.4
.2
.0
SD ln(RT)
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
Number of Examinees

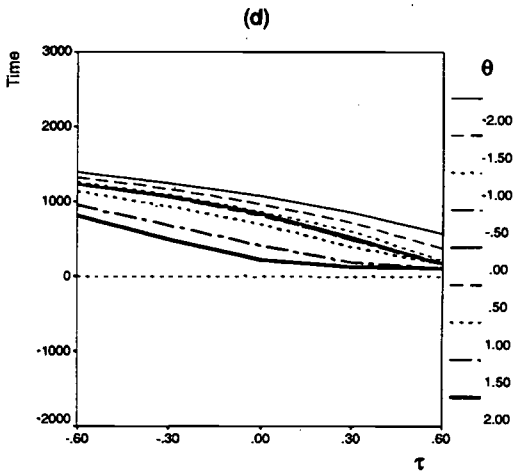












Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-98-06 W.J. van der Linden, D.J. Scrams & D.L.Schnipke, *Using Response-Time Constraints in Item Selection to Control for Differential Speededness in Computerized Adaptive Testing*

RR-98-05 W.J. van der Linden, *Optimal Assembly of Educational and Psychological Tests, with a Bibliography*

RR-98-04 C.A.W. Glas, *Modification Indices for the 2-PL and the Nominal Response Model*

RR-98-03 C.A.W. Glas, *Quality Control of On-line Calibration in Computerized Assessment*

RR-98-02 R.R. Meijer & E.M.L.A. van Krimpen-Stoop, *Simulating the Null Distribution of Person-Fit Statistics for Conventional and Adaptive Tests*

RR-98-01 C.A.W. Glas, R.R. Meijer, E.M.L.A. van Krimpen-Stoop, *Statistical Tests for Person Misfit in Computerized Adaptive Testing*

RR-97-07 H.J. Vos, *A Minimax Sequential Procedure in the Context of Computerized Adaptive Mastery Testing*

RR-97-06 H.J. Vos, *Applications of Bayesian Decision Theory to Sequential Mastery Testing*

RR-97-05 W.J. van der Linden & Richard M. Luecht, *Observed-Score Equating as a Test Assembly Problem*

RR-97-04 W.J. van der Linden & J.J. Adema, *Simultaneous Assembly of Multiple Test Forms*

RR-97-03 W.J. van der Linden, *Multidimensional Adaptive Testing with a Minimum Error-Variance Criterion*

RR-97-02 W.J. van der Linden, *A Procedure for Empirical Initialization of Adaptive Testing Algorithms*

RR-97-01 W.J. van der Linden & Lynda M. Reese, *A Model for Optimal Constrained Adaptive Testing*

RR-96-04 C.A.W. Glas & A.A. Béguin, *Appropriateness of IRT Observed Score Equating*

RR-96-03 C.A.W. Glas, *Testing the Generalized Partial Credit Model*

RR-96-02 C.A.W. Glas, *Detection of Differential Item Functioning using Lagrange Multiplier Tests*

RR-96-01 W.J. van der Linden, *Bayesian Item Selection Criteria for Adaptive Testing*

RR-95-03 W.J. van der Linden, *Assembling Tests for the Measurement of Multiple Abilities*

RR-95-02 W.J. van der Linden, *Stochastic Order in Dichotomous Item Response Models for Fixed Tests, Adaptive Tests, or Multiple Abilities*

RR-95-01 W.J. van der Linden, *Some decision theory for course placement*

RR-94-17 H.J. Vos, *A compensatory model for simultaneously setting cutting scores for selection-placement-mastery decisions*

RR-94-16 H.J. Vos, *Applications of Bayesian decision theory to intelligent tutoring systems*

RR-94-15 H.J. Vos, *An intelligent tutoring system for classifying students into Instructional treatments with mastery scores*

RR-94-13 W.J.J. Veerkamp & M.P.F. Berger, *A simple and fast item selection procedure for adaptive testing*

RR-94-12 R.R. Meijer, *Nonparametric and group-based person-fit statistics: A validity study and an empirical example*

RR-94-10 W.J. van der Linden & M.A. Zwarts, *Robustness of judgments in evaluation research*

RR-94-9 L.M.W. Akkermans, *Monte Carlo estimation of the conditional Rasch model*

RR-94-8 R.R. Meijer & K. Sijtsma, *Detection of aberrant item score patterns: A review of recent developments*

RR-94-7 W.J. van der Linden & R.M. Luecht, *An optimization model for test assembly to match observed-score distributions*

RR-94-6 W.J.J. Veerkamp & M.P.F. Berger, *Some new item selection criteria for adaptive testing*

RR-94-5 R.R. Meijer, K. Sijtsma & I.W. Molenaar, *Reliability estimation for single dichotomous items*

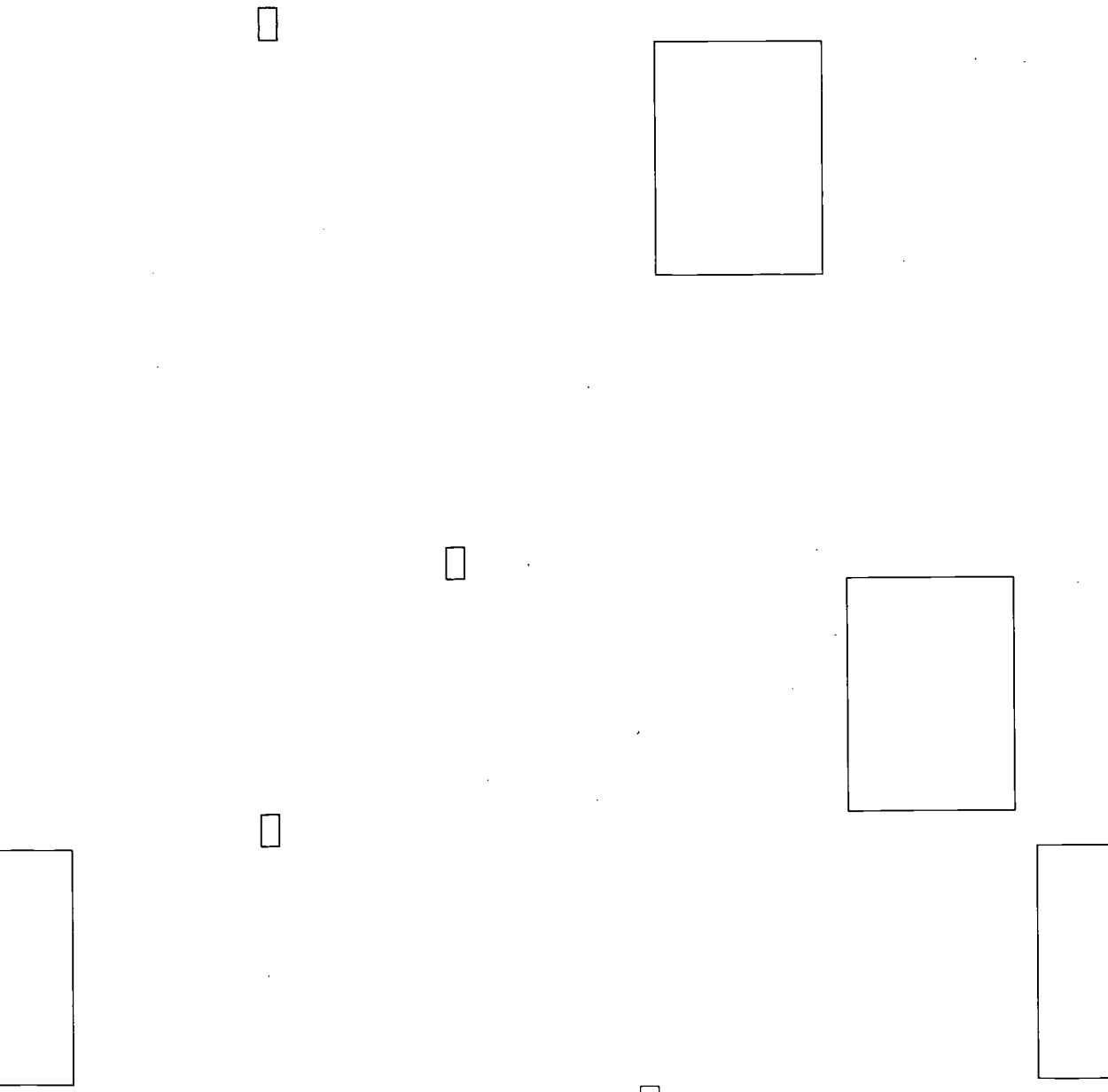
RR-94-4 M.P.F. Berger & W.J.J. Veerkamp, *A review of selection methods for optimal design*

RR-94-3 W.J. van der Linden, *A conceptual analysis of standard setting in large-scale assessments*

RR-94-2 W.J. van der Linden & H.J. Vos, *A compensatory approach to optimal selection with mastery scores*

RR-94-1 R.R. Meijer, *The influence of the presence of deviant item score patterns on the power of a person-fit statistic*

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, Mr. J.M.J. Nelissen, P.O. Box 217, 7500 AE Enschede, The Netherlands.



faculty of

EDUCATIONAL SCIENCE AND TECHNOLOGY

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☑ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").